210A Week 6 Notes

A note on n and statistical testing

This week we're going to discuss significance tests. It's worth stressing that this is statistical significance, which oughtn't be overly confused with social significance or theoretical significance. It's increasingly common to take statistical significance with a grain of salt for the excellent reason that statistical significance is based on standard error and standard error is in turn based on n. So just as God is on the side of the big battalions, significance is on the side of the big datasets.

Generally speaking, if something is statistically insignificant with a small sample it still *might* be socially significant. If you have decent n, then they roughly correspond. If you have huge n then statistical significance is really a necessary but not sufficient condition for social significance. Of course, this begs the question of how do you tell social significance if you can't just equate it with p? The answer is nothing more nor less than the irreducibly subjective practice of you look a the size of the parameter and ask yourself, does this *seem* important? Could I convince my colleagues and peer reviewers that it is?

The null hypothesis as a confidence interval threshhold

Technically statistical significance is all about hypothesis testing. However statistics uses "hypothesis" in a very weak sense to mean any kind of prediction, whether or not that prediction has a huge lit review. Usually the "hypothesis" is nothing more elaborate than "the amount here is more than nothing" or "this is different from that."

Most of the time the alternative hypothesis, or H_a, is just "this is different," and the null hypothesis, or H_0, is just "this is the same." These definitions of H_a and H_0 are built into most of our software by default. Usually the null hypothesis is a point or threshhold of some kind and you're trying to see whether the confidence interval of the estimate encompasses this point or threshhold. The book gives a lot of examples where the threshhold is some arbitrary value. However software usually assumes this threshhold is zero so in regression output the question tested by significance testing is "does the coefficient equal zero?" If you want some arbitrary threshhold other than zero, it's usually easier to recenter your data on zero (by subtracting the threshhold) than it is to fight with the software.

Last week I gave the example of a point estimate for Obama's vote as 0.54 with a se of .015. To the extent that we just care about whether the poll projected him to win and we don't care by what margin, we asked what is the probability that he'll win (by any margin)? This week we will go over that again, but more formally.

We can phrase this as a statistical significance test. What we're really interested in is whether $obamavote > .50$. Note that we don't just care whether it's *different* from .50 (which would be a two-tailed test) but whether it's *bigger*

than .50 (1 tail). Another way to think of it is that the .54 estimate could be off in two ways, he could lose or he could win a landslide. Since a landslide is still a win we don't care about the right-tail and we're only interested in the left-tail (he loses). To put it formally:

$H_0 : obamavote \leq .5$

$H_a : obamavote > .5$

Remember that H_0 is basically the thing you're trying to disprove and H_a is what you're trying to prove. First, it's often easy to think of H_0 as zero, if for no other reason than the computer likes it this way. So let's make it so by subtracting our threshhold of .50 from H_a, H_0, and our parameter estimate. Conceptually we can think of this as going from thinking about obama-vote to obama-victorymargin where a negative obama-victorymargin is a mccain win.

So now $H_0 : obamavictorymargin \leq 0$ and our point estimate of *obamavictorymargin* as .04. Basically, our H_0 is that Obama loses, but our best guess is that he'll win by four points.

Second, let's figure out the p-value, or what is probability that Obama lost. To do this we have to figure out how much bigger our estimate (.04) is than 0 in terms of *se*.

$t = (.04 - 0)/1.5 = 2.7$

Now we look at the table and see how much density is in one tail at least 2.7 sigmas out. Because we have a reasonably large n we can just use the normal distribution as a reasonable approximation of the t distribution. We get p=0.0035. So there was only a 0.35% that Obama would lose (assuming we trust our data). We'll pick apart what else exactly *"p"* means later. Another way to put this is that we have a 99.65% confidence level that the confidence interval does not include a McCain win.

That was a one-tailed test (specifically, the left tail), but two-tailed tests are more common. Imagine that you're a network news anchor and you're trying to figure out if you can buy nonrefundable plane tickets for the day after the election. In this sense, you don't care who wins, you just want to know if it's going to be a nail-biter like 2000 such that it remains uncalled for a couple weeks and you have to give up your nonrefundable tickets to stay on the air and talk about lawsuits and hanging chads.

So what we want to know is the probability that the vote is exactly .50, and we don't care if McCain wins clearly or Obama wins clearly, so long as somebody wins and we get to go on vacation. For this we do a two-tailed test where $H_0 : obamavote = .5$ and $H_a : obamavote \neq .5$. Again, it's easiest to center the variable and prediction so that zero means "a tie." We now ask about the probability that there could really be a tie.

The way we do this is we go back to the $CI + CL$ business. Treat your point estimate as drawn from a distribution and you ask how wide do you have to make the CI so that it just barely encompasses the threshhold (which you should set to zero). In practice, this means you do a one-tailed test, but then you double the density.

$t = .04/ \pm 1.5 = \pm 2.7$

If you add the two-tails together, the density at ± 2.7 is double that at just 2.7. Hence we get $p = 2(0.0035) = 0.007$. The newscaster with vacation plans is kind of a silly example. Mostly you'd want to do this when the question is whether there is *something* going on. Is it *different* or is it *distinguishable*? You can ask this question without asking in what direction does it differ.

This brings up the issue of choosing one-tail vs two-tail. Two-tailed means no expectation whereas one-tailed includes a more precise prediction. Thus if you take Popper seriously one-tailed hypotheses seem more scientific. Unfortunately they tend to be a bad idea in practice.

1. One-tailed tests are more sensitive

Notice that p of two-tails is always double p of one-tail. Since we like having firm conclusions (high CL or low α) this means that two-tailed tests get you less bang for the buck than one-tailed tests. In practice, using one-tailed can be cheating to basically mislabel a result where $p<.10$ to one where $p<.05$. Unless you have firmly committed yourself to a very clear one-tailed hypothesis in advance of the analysis, using a one-tailed test is sketchy.

2. You should have some humility as your data can surprise you.

Take the Sarah Silverman hypothesis, that old Jews are less likely to vote for Obama than young Jews. Say we think this is such a good hypothesis that we do a one-tailed test and we find that this is not true. Can we then conclude that there's no political difference between old and young Jews? No, we can only conclude that young Jews are not *more* likely to support Obama than old Jews. If we did a two-tailed test, we'd find that there is a difference but it's the *opposite* of what Silverman expected, actually *young* Jews are most likely to vote for McCain.

The bottom line is that by default you should use a two-tailed test and treat one-tailed tests very carefully.

Note that the silverman example is comparing the mean of two populations (ie, Jewish McCain voters and Jewish Obama voters, compared by mean age). This is a t-test of means (or simply the "t-test"), which you use when you have one binary variable and one continuous variables. We'll cover this next week. The week after that we'll do the χ^2, which you use for two categorical variables.

Confidence Level and the Trade-Off Between Type I and Type II Error

One simple way to think of "p" is as basically how worried you ought to be that your null is true. Today we'll complicate that. Statisticians like to use a lot of double negatives so they can avoid ever saying anything, but for I'll fudge the math slightly for the sake of the English language. In a nutshell, there are three things you can say

1. H_a is probably true
2. H_0 is probably true
3. beats me

The problem is that p conflates the last two. As such, *technically* a high p should not be interpreted as "reject H_a," but only as "do not reject H_0." It's a lot like a verdict of "not guilty" in court. We say "not guilty" instead of "innocent" because a jury may think that somebody is *probably* guilty but not beyond a reasonable doubt. This fuzziness brings in the issue of our standards or what is "beyond a reasonable doubt" and what is "statistically significant"?

For instance, did R Kelly do obscene things with an underage girl, or was the videotape showing this actually an elaborate forgery created by a conspiracy of disgruntled former associates and con men? It's certainly *conceivable* that the tape was a forgery, but is this a "reasonable" doubt?

This where we get into the cliche about "is it better that ten guilty men go free than one innocent man be punished?" OK, but what about twenty guilty men? or 100? It's a lot like Genesis 18, with Abraham trying to talk God out of destroying Sodom in a rain of hellfire. The basic issue is that if your standards are too credulous you will convict some of the innocent (but probably get most of the guilty). On the other hand if, like the R. Kelly jury, your standards are ridiculous then you will will free the innocent (but probably not convict the guilty).

In law they have crude threshholds like "preponderance of evidence" or "reasonable doubt" for evidence and "rational basis" or "strict scrutiny" for case law. In statistics we have alpha, which is the threshhold of p you're willing to accept as basically deciding for H_0 or H_a. By convention, we treat .05 as the default alpha. If $p<.05$, call it "statistically significant" and put a "*" next to it. Several other common threshholds are .10, .01, and .001 but these all have to be specified as just plain "*" or "significance" means .05.

When we call something "significant" we're saying H_a is probably true and H_0 false. Alpha is how often we're willing to be wrong about H_0 probably being false. This is called type I error or "false positive"– how often we should have accepted the null but didn't. By demanding a lower (that is stricter) alpha we can drive down type I error so we're less likely to reject H_0 when it's actually true.

The problem is that by being so demanding we may reject some good theories. This is type II error – accepting H_0 when we should have rejected it. Page 160 of the book has a chart, which we can simplify by eliminating the CYA double negatives to be:

	we accept H_a	we accept H_0
	(common with high alpha)	(common with low alpha)
H_0 actually true	Type I error	valid
H_a actually true	valid	Type II error

We'd like our conclusions to always be valid but there is a trade off. A high alpha (i.e., loose standards of evidence) implies a lot of Type I error. A low alpha (i.e., high standards of evidence) implies a lot of Type II error. In some situations you may be worried about different types of error. So if H_a is finding the defendant at fault, common law justice worries more about Type I error than Type II error is criminal law (hence the "reasonable doubt" standard) but

worries about them equally in tort law (hence the "preponderance of evidence" standard).

I like to think of it as making popcorn in the microwave. What we'd like is for all of the popcorn to pop and none of it to burn, but this is impossible. If you worry too much about popping every kernel, some gets burnt and if you worry too much about some getting burnt then some remain unpopped. Nonetheless, you can usually get *most* of the bag popped but not burnt – unless you have a bad microwave or really stale popcorn kernels. Think of having a high standard error (usually because of a small n) as like having stale popcorn kernels.

One of the biggest complaints about President Bush was that he ignored the (prescient) August 6, 2001 "presidents daily briefing" titled "Bin Ladin Determined To Strike in US." *aha! W knew.* However the president gets lots of warnings (most of which are false alarms), which is one of the reasons that his response to the PDB was not "tell Panetta to shut down all air travel immediately" but "All right. You've covered your ass, now." After 9/11 they effectively raised their alpha, which is how we got the permanent "orange" alert and the Iraq War. Failing to prevent 9/11 was Type II error and expecting to find Iraqi WMD was Type I error. It's impossible to quantify, but it seems like intelligence has a lot of standard error so you can see how we get such problems.

In quant research we face similar issue. The stakes are usually lower than war but higher than popcorn. If you make your confidence level too low, you require too narrow of a confidence interval and you'll miss out on findings, but if you have a confidence interval that's too wide you'll get a lot of spurious findings.

So what's the solution? It's never going to be perfect but more information reduces the trade-off. In criminal justice we can adequately fund detectives and forensic labs. In counter-terrorism we can get more intelligence.

In science can get more n. The size of uncertain region is function of standard error which in turn is a function of n. So the type I & type II trade off especially severe for small studies, but not so much for huge datasets.

Publication Bias

It's hard to get things in journals and generally reviewers don't like results that are inconclusive or hard to interpret. Therefore when you get such results, you tend to give up or, if you persist, find yourself unable to publish them. Note that with an alpha of .05 we would expect almost one in 20 completely random tests to be statistically significant. However, if only significant results get published then it's actually much worse than that since the null findings never make it into the literature and thus are censored. In effect, only the "accept H_a" column of the above table appears in the literature.

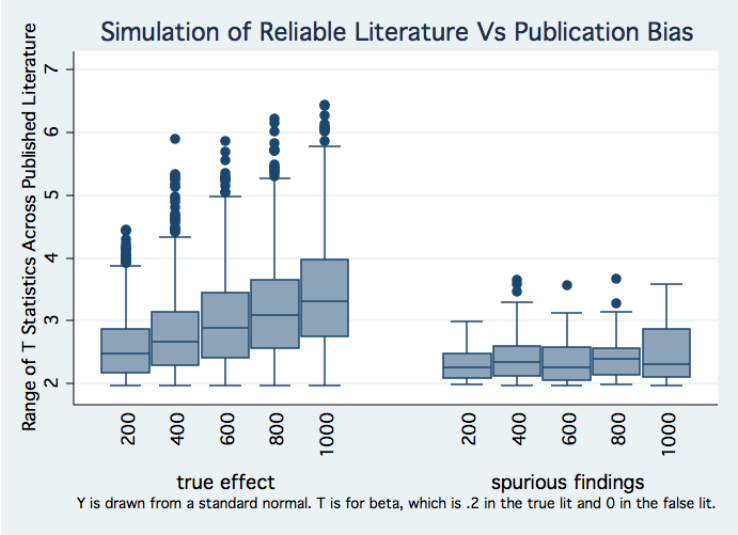
The implication is that while Type I error should only occur one of twenty times for any particular study, it should be much more common in that subset of research that is actually published – a methodological headache even as it's also a nice example of the theoretical issue of macro phenomena being *emergent* from micro dynamics. Thus you can't really trust the alpha as stated in the

literature because publication bias censors findings of the null. Fortunately, meta-analysis let's you tell whether overall the literature is accurate, though it can't tell you whether any one study is accurate.

If we had an accurate theory, most tests of it would be significant and would get published so we would have little censorship. Without censorship, we would expect them all to find about the same effects and thus the significance would rise as a *positive* function of the square root of degrees of freedom. So with good literature $\beta \neq f(df)$ and $t = f(df)$.

On the other hand, if there were no true association, only those studies which randomly appeared to show an association would get published and those which (accurately) showed no association would be censored. Since the censorship is at a threshhold of t>1.96, and$t = \beta/se$ and $se = f(\sqrt{df})$ this implies that the smaller the *df*, the larger the parameter needs to be before we can see it. Therefore a censored literature will show $\beta = f(df)$ but $t \neq f(df)$.

I used random data to run a simulation of two models, each of which has a binary variable predicting a continuous outcome. (See the do-file here). Only significant results are published. Where there actually is a true effect, t increases with sample size. Where there is no true effect, t tends to stay the same regardless of n.



One of the best empirical examples of this in social science is the debate over the minimum wage. (Last time I talked about this I confused a few of the details but I've since checked everything). One of the basic axioms of economics is that demand curves slope downwards (raise the price and you decrease the quantity demanded) which in the context of labor implies that if you raise the minimum wage you'll decrease the quantity of demand for unskilled labor (i.e., raise unemloyment). A series of studies had indeed demonstrated this.

Card and Krueger did a meta-analysis of this literature and found that

published studies with big n have small β and t hovering right about 2. The scatterplot labeled "Card and Krueger 1995" shows their meta-analysis, which shows a flat relationship between sample size and statistical significance, a clear statistical signature for a bullshit literature.[1]

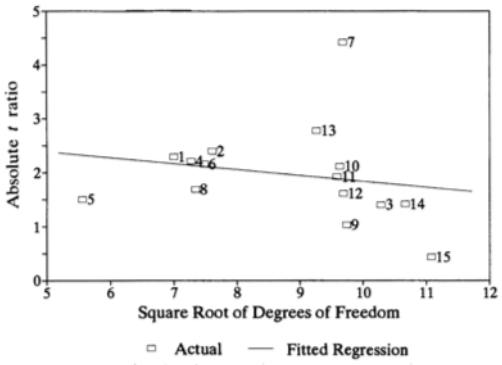


FIGURE 1. RELATION OF ESTIMATED t RATIOS TO SAMPLE SIZE

Source: Card, David and Alan B. Krueger. 1995. "Time-Series Minimum-Wage Studies: A Meta-analysis" *The American Economic Review* 85:238-243

This meta-analysis was necessary in order to explain the discrepancy between the literature and their finding in another paper that a minimum wage hike in New Jersey from $4.25 to $5.05 actually *increased* employment at fast food restaraunts.[2]

The fast food study is often interpreted as evidence that the minimum wage is actually good for unskilled employment, but there is no good theoretical reason to suppose this and many of the proffered reasons (efficiency wage, monopsony) don't actually make theoretical sense in this context. Rather the most plausible interpretation is that in the short-run the demand for unskilled labor is inelastic to small changes in wage and since there is no true effect we will every so often randomly get results that falsely imply that small minimum wage increases the demand for labor (Card and Krueger 1994) or decrease the demand for labor (everybody else who has published on the issue). Note that Card and Krueger have not spoken to the long-run effects, nor to effects of very large minimum wage boosts ("living wage"), but neither has anyone else proven that there will be an effect under these broader conditions. We simply have to say that we lack good empirical evidence for the effects of minimum wage outside of Card and Krueger's scope conditions, although we can speculate from theory.

Bottom line, censorship implies that alpha is actually a lowball estimate of Type I error and you should be especially suspicious of a literature that is

[1] Card and Krueger's explanation for why the curve is slightly negative rather than completely flat is that it's easier to publish null findings with a large dataset.

[2] Card, David and Alan B. Krueger. 1994. "Minimum Wages and Employment: A Case Study of the Fast-Food Industry in New Jersey and Pennsylvania" *The American Economic Review* 84:772-793.

mostly based on small n and those few large n studies that it contains have much smaller effects than the small n studies.